LoftQ: LoRA-Fine-Tuning-Aware Quantization for Large Language Models

Year: 2023 | *Citations: 190* | *Authors: Yixiao Li, Yifan Yu, Chen Liang*

Abstract

Quantization is an indispensable technique for serving Large Language Models (LLMs) and has recently found its way into LoRA fine-tuning. In this work we focus on the scenario where quantization and LoRA fine-tuning are applied together on a pre-trained model. In such cases it is common to observe a consistent gap in the performance on downstream tasks between full fine-tuning and quantization plus LoRA fine-tuning approach. In response, we propose LoftQ (LoRA-Fine-Tuning-aware Quantization), a novel quantization framework that simultaneously quantizes an LLM and finds a proper low-rank initialization for LoRA fine-tuning. Such an initialization alleviates the discrepancy between the quantized and full-precision model and significantly improves generalization in downstream tasks. We evaluate our method on natural language understanding, question answering, summarization, and natural language generation tasks. Experiments show that our method is highly effective and outperforms existing quantization methods, especially in the challenging 2-bit and 2/4-bit mixed precision regimes. The code is available on https://github.com/yxli2123/LoftQ.